A Conclusive Methodology for Rating OCR Performance

Nathan E. Brener, S.S. Iyengar, and O.S. Pianykh
Department of Computer Science, Louisiana State University, Baton Rouge, LA 70803.
E-mail: brener@bit.csc.lsu.edu

One of the most challenging topics in the automatic document rating process is the development of a rating scheme for the image quality of documents. As part of the Department of Energy (DOE) document declassification program, we have developed a generalized rating system to predict the optical character recognition (OCR) accuracy level that is achieved when processing a document. The need for such a system emerged from the declassification of degraded, typewriter-era documents, which is currently a time-consuming manual process. This article presents the statistical analysis of the most influential document quality features affecting OCR accuracy, develops consistent predictive models for four currently used OCR engines, and studies the applicability of different OCR products to the DOE document declassification process. This study is expected to lead to an efficient and completely automated document declassification system.

Introduction

Efficient, reliable, and fully automated optical character recognition (OCR) has become one of the most important problems in modern document analysis. OCR is a method of transforming a page image into a text file. The goal of this transformation is to make letters, words, and symbols printed on a page identifiable. *Document rating systems* attempt to rank page images on the basis of the degree to which they can be accurately transformed into text by OCR.

Although modern OCR engines are implemented with fast pattern recognition techniques, document rating still requires human interaction. This interaction becomes especially unavoidable in the rating of degraded text documents, where document quality is substantially reduced by age, copying, faxing, low-resolution printing, typing, and other processes. Because rating such documents manually is a tedious and expensive task, a fast and fully automated rating system would be highly beneficial in the document rating process.

Received June 5, 2001; revised August 30, 2004; accepted August 30, 2004

 • Published online 26 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20214

This article introduces an efficient automated document rating scheme that is accurate and does not require human assistance. In the first part, we develop a minimal set of document quality parameters sufficient for high-fidelity rating. These parameters are then used to predict the actual OCR accuracy levels obtained with four different OCR products for a set of old typewritten pages. Finally, we evaluate the four OCR engines with respect to a population of degraded text images and find the engine that gives the best performance on these images (Senior & Robinson, 1998; Cannon, Kelly, Iyengar, & Brener, 1997; Tang, Tu, Lee, Lin, & Shyu, 1998).

Department of Energy Declassification Problem

In our development of an automated document rating system, we used a set of low-quality typewriter-era documents provided by the Department of Energy (DOE) Office of Declassification (OD). Currently, OD has approximately 230 million pages of documents waiting to be declassified. At present, these documents are being declassified manually, a time-consuming and labor-intensive task. Clearly, this declassification process needs to be automated, with the first step being the conversion of the documents to text files using OCR engines. In order for the automated system to run successfully, the OCR conversion must be done at a high level of accuracy. Current OCR commercial products are designed primarily for documents produced by laser printers and typically achieve high levels of accuracy on such documents. However, much of the OD document population was produced by typewriters in the 1940s and 1950s and is in sharp contrast to laser printer documents. As a result, many of the modern OCR technologies may not achieve sufficiently high accuracy on the OD typewriter-era documents. Thus we need an automated system that can evaluate a document's readability with different OCR engines and predict the OCR accuracy rate for each engine. In this way, we can select the best OCR engine to use on each document.

In order to develop such an automated system, we need to determine the image quality factors that cause errors for a particular OCR engine, develop an efficient tool for predicting

OCR accuracy rates (Cannon, Kelly, Iyengar, & Brener, 1997), and find the most appropriate OCR product for processing typewriter-era documents. These aims were accomplished by performing a detailed statistical analysis of the effect of different image quality parameters on OCR accuracy. This analysis is expected to lead to an automated system that will substantially improve the OD declassification process.

Automated Document Rating Scheme

Extracting Image Quality Parameters

In order to determine which factors affect OCR accuracy, we developed a 131-page test suite selected from approximately 550 pages provided by Dyncorp (Crandall & Amsler, 1996). All pages were typewritten and were chosen to contain various levels of text degradation, representing the lower-quality documents that are part of the OD document population waiting to be declassified. As a first step, the pages were scanned and digitized as binary images in compressed tagged image file (TIF) format; each image was approximately 2000 by 3000 pixels.[1] Then the four leading commercial OCR engines were used to process the 131 images; we will refer to these engines as A, B, C, and D. For each page, the actual OCR accuracy rate was manually calculated on a [0, 100] scale as the percentage of correctly identified characters, resulting in a 131-dimensional vector of accuracy rates for each OCR engine.

Our goals were to eliminate the tedious manual ranking and to develop a simple but efficient image-processing tool to predict the OCR accuracy rate from morphological parameters of a scanned page. For increased efficiency, only minimal image processing was applied. First, the top and bottom of each text line were located from the vertical black pixel density histogram, as described in Senior and Robinson (1998). Then the beginning and the end of each text line were determined from the horizontal line density histograms. This process enabled us to identify correctly the location of nearly all text lines, including titles, headers, and, in some cases, page numbers. After these text-containing rectangular regions were located, the bottom, top, left, and right margins were removed from the analysis, and the remaining image was subdivided into text-containing and complementary between-line regions, as shown in Figure 1. Average text line height was chosen as a character height estimate *CH*. From the horizontal line density histogram each text-containing rectangle was divided into regions with above-average horizontal density, corresponding to the characters (sometimes broken or touching), and the average character width *CW* was estimated. Once the character-containing regions were filtered out from the between-character noise, the histograms for black connected[2] components *CC*, black connected components within text line regions *LCC*, white connected components within text line regions *WLCC*, and "thick" black connected components[3] within text line regions *TCC* were determined. For each statistic, its mean and variance were computed to estimate both the most probable parameter value and the amount of randomness it can have (Table 1).

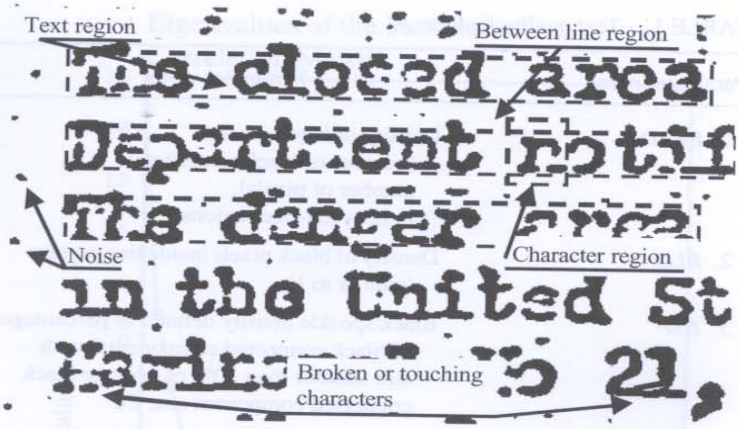


FIG. 1. Text image segmentation.

Document Quality Multifeatures

The OCR predictability analysis resulting from the document segmentation described was based on a natural assumption that the character-based OCR recognition accuracy rate will be proportional to the quality of the text-containing image.[4] Therefore, we assumed that given the OCR rate R and a set of image quality parameters p_i, $i = 1, \ldots, n$, one can efficiently approximate R with a linear regression model:

$$R = b_0 + \sum_{i=1}^{n} b_i p_i + e,$$

where e, the regression error, is minimized with an appropriate choice of regression coefficients b_i (Iyengar & Rao, 1983). However, we did not assume any a priori knowledge of the specific set of predictor features p_i. Therefore, we initially computed $n = 36$ text quality parameters, which are described in Table 1 and illustrated in Figure 2.

These parameters were grouped into seven generic classes shown in Figure 3. The parameters within each class tend to be correlated; for instance, large character height *CH* is likely to correspond to a large font size, resulting in a large character width *CW*. Because the initial number of parameters $n = 36$ was large, optimal parameter reduction was applied, as explained in the next section.

[1]For better analysis, all pages were scanned with 3×-4× magnification.
[2]We used 8-connectivity.
[3]Same as LCC but only for the black pixels with all eight black neighbors.
[4]This is less true for the word-based OCR, which would require an additional and complex linguistic analysis, avoided here for efficiency. Moreover, we observed the high correlation between character-based and word-based OCR rates.

TABLE 1. Text quality features p_i.

Parameter name	Parameter value	Parameter name	Parameter value
1. *BBLD*	Density of black pixels, [number of black pixels/total number of pixels], in between-line regions	10. *LCCD*	Average line connected component density[b]
2. *BLD*	Density of black pixels inside line regions (similar to 1)	11. *RSS*	Percentage (with respect to total number of line connected components) of line connected components with width and height both below 0.5*LCCH*
3. *BSP*	Black speckle density defined as percentage of black connected components[a] with size smaller than 20% of average black connected component size *CC*	12. *RLS*	Percentage (see 11) of line connected components with width above and height below 0.5*LCCH*
4. *CC, CCV*	Average black connected component size and its variance	13. *RSL*	Percentage (see 11) of line connected components with width below and height above 0.5*LCCH*
5. *CCH, CCHV*	Average black connected component height and its variance (standard deviation)	14. *RLL*	Percentage (see 11) of line connected components with width and height both above 0.5*LCCH*
6. *CCW, CCWV*	Average black connected component width and its variance	15. *TCC, TCCV, TCCH, TCCHV, TCCW, TCCWV*	"Thick" line connected component statistics, computed similarly to the overall connected component parameters in 4–6, but now only for thick connected components inside text line regions. Thick connected component includes only black pixels with all black 8-pixel neighborhood[c]
7. *CH, CHV*	Average character height (as average text line height) and its variance		
8. *CW, CWV*	Average character width and its variance found from horizontal line density histogram		
9. *LCC, LCCV, LCCH, LCCHV, LCCW, LCCWV*	Line connected component statistics, computed similarly to overall connected component parameters in 4–6, but now only for connected components inside text line regions.	16. *WLCC, WLCCV, WLCCH, WLCCHV, WLCCW, WLCCWV*	White line connected component statistics, computed similarly to the line connected component parameters in 9, but for the white pixel connected components
		17. *WSP*	White speckle density, computed similarly to BSP, but for white pixels

[a]A black (white) pixel is defined as a connected component pixel if all its eight neighboring pixels are black (white). A black (white) connected component is a connected set of connected component pixels of the same color (either black or white).

[b]Found as the average of the ratio (number of connected component pixels/[connected component width × connected component height]).

[c]That is, all eight neighbors of the pixel must also be connected component pixels of the same color (either black or white). This statistic was defined to eliminate small connected components introduced by random noise.

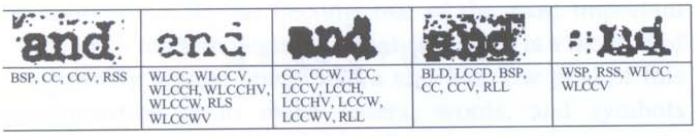


FIG. 2. OCR features for degraded characters.

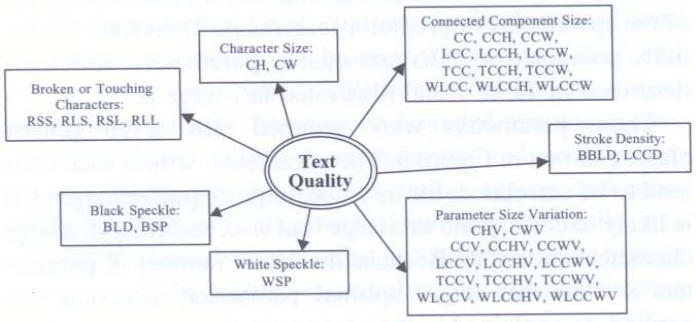


FIG. 3. Text multifeature set.

Reducing the Number of Predictor Variables

Similarly to the OCR engines, each parameter p_i in our analysis was represented by a 131-dimensional vector, consisting of the parameter value computed for all 131 test pages. The best two-dimensional parameter space representation based on interparameter Euclidean distances is shown in Figure 4 (left). The figure was obtained by mapping parameter representation vectors from their original 131-dimensional space to a 2-dimensional space, with the mapping chosen to preserve the relative distances between the parameters optimally. This plot approximately visualized the proximity of parameters and suggested the need for a reduced parameter set because many features in Figure 4 (left) nearly overlap. Therefore, we used the SAS® predictor selection tool to choose the best (corresponding to the lowest possible mean square error [MSE]) set of predictor variables for each model size n and each OCR engine as follows: First, the optimal number of significant parameters was estimated with principal component analysis. For the given 131-dimensional 36-parameter vector set we determined that approximately the first[5] five to six eigenvalues of the parameter correlation matrix

$$C = \left[C_{i,j} = (p_i, p_j) = \sum_{d=1}^{131} p_i^{(d)} p_j^{(d)} \right]_{i,j=1}^{n=36}$$

[5]Ordered by their magnitude.

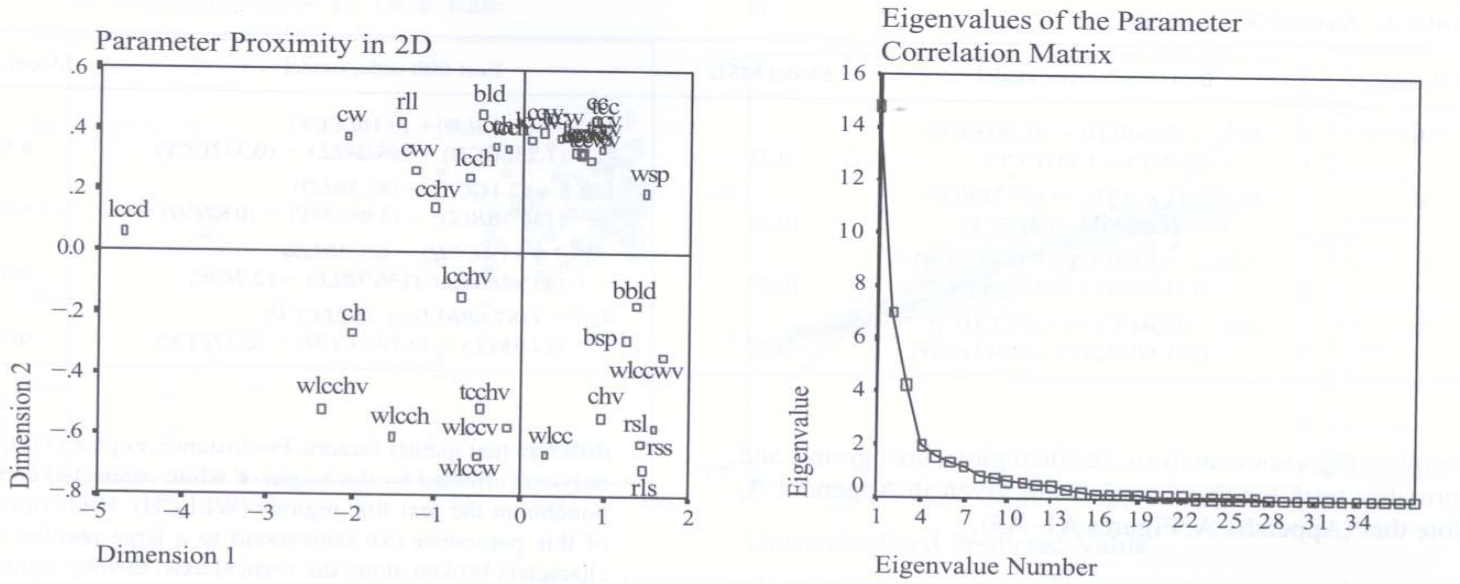


FIG. 4. Document quality parameters.

account for the majority (82%) of the total parameter set variance. Figure 4 (right) represents the contribution of each eigenvalue to the total parameter set variance. Therefore, principal component analysis suggested that the dimensionality of the parameter space could be reduced to five to six dimensions. This dimensionality estimate allowed development of a compact OCR rate-predicting scheme based on linear regression with four to six predictor variables as presented below.

To eliminate correlated parameters, all four OCR engines, A, B, C, and D, and the averaged OCR values[6] were considered for the parameter reduction analysis. For each linear regression model of size n and each OCR engine, the 131-dimensional OCR rate vector R was regressed on every possible n-variable subset, $1 \le n \le 36$, of the complete parameter set, and the regression with minimum MSE was determined. We started this analysis with the averaged OCR rate and found that four text quality factors, namely, *CCH*, *BBLD*, *LCCV,* and *TCCV,* were sufficient to achieve a remarkably high 0.9 model-to-response correlation, and mean square error MSE as low[7] as 8.41 (with response OCR values varying in the [0, 100] range). This best four-predictor model for the OCR accuracy rate, denoted by *OCR,* was computed as follows:

$$OCR = 60.66 + (1.34CCH) - (129.61BBLD) + (0.21LCCV) - (0.36TCCV)$$

The best five-predictor model was determined to be

$$OCR = 50.03 + (1.12CCH) - (123.95BBLD) + (0.22LCCV) - (0.38TCCV) + (18.35RLL)$$

and led to an MSE of 8.10.

[6]Determined as follows: For each page, the smallest and the highest OCR engine rates were dropped, and the remaining two averaged.

[7]Full model with all 36 predictors gives 0.96 correlation and 6.45 error.

Then a similar model selection procedure was applied to the A, B, C, and D engines separately. For each OCR engine and each number of features $n = 1$ to 36, the optimal linear regression model of size n was found, and the respective mean square error was determined. The resulting decrease in minimal MSE value, with respect to the linear regression model of size n, is shown in Figure 5. As one can observe, increasing the number of predictor variables beyond six resulted in a more moderate MSE decrease compared to the smaller model sizes.

Therefore, we limited model sizes n to 4, 5, and 6 and determined optimal variable sets for each OCR engine. This method permitted us to reduce regression MSE values below 9.5 (for size $n = 6$), a substantial 60% improvement compared to the results obtained with another OCR prediction method described in Cannon and associates (1997). Statistical analysis for these models is presented in the following section.

Linear Prediction for OCR Engines

The optimal fourth- and fifth-order regression results for the four OCR engines are summarized in Table 2, and the

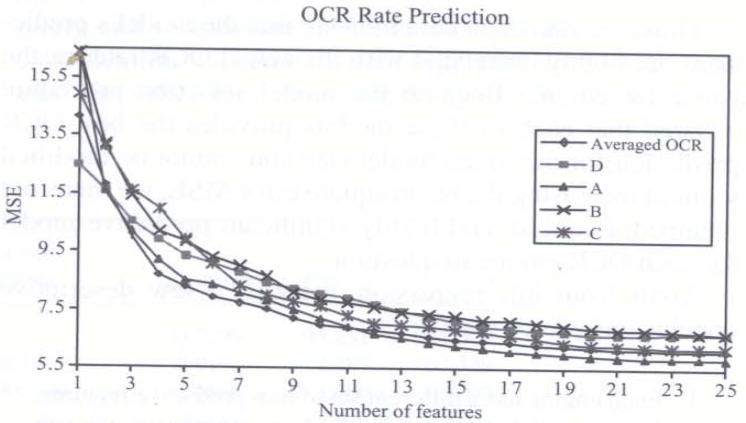


FIG. 5. Predicting OCR rates.

TABLE 2. Optimal OCR predictors.

OCR engine	Best fourth-order model	Model MSE	Best fifth-order model	Model MSE
A	$56.0 - (84.8BLD) + (0.18LCCV) + (66RLL) - (.34TCCV)$	9.21	$46.2 - (68.2BLD) + (0.16LCCV) + (1.23LCCH) + (48.0RLL) - (0.33TCCV)$	8.52
B	$54.2 + (1.8CCH) - (167.7BBLD) + (0.2LCCV) - (0.4TCCV)$	10.21	$122.7 + (2.1CCH) - (82.3BLD) - (144.4BBLD) - (3.08CWV) - (0.87CH)$	9.84
C	$-50.2 - (0.04CCV) + (1.9LCCH) + (123.8RSS) + (111.1RLL)$	10.99	$-13.7 + (1.4CCH) - (81.2BLD) + (135.7RSS) + (136.3RLL) - (2.7CW)$	9.93
D	$102 - (0.04CC) + (1.67CCH) - (201.5BBLD) - (3.0WLCCH)$	9.92	$92.7 - (187.6BBLD) + (0.2LCCV) + (24.5RLL) - (2.7WLCCH) - (0.3TCCV)$	9.36

complete regression analysis, residual plots, histograms, and normality tests for these models are given in Appendix A. Note that (Appendix A, Figures A1–A5):

1. All regression models are significant: the analysis of variance (ANOVA) significance level, which indicates the degree to which these models are inappropriate, was found to be below 0.0001 in all cases, which indicates that all four OCR engines and their average are strongly influenced by the values of their respective five-feature sets.
2. All feature coefficients b_i shown in the respective coefficients tables are also significant on at least the 0.0001 level. Therefore, each feature in the predictive set has a strong influence on the respective OCR rate.
3. The mean square prediction error for OCR engines ranges from $\sqrt{65.614} = 8.1$ for the averaged OCR values to $\sqrt{98.601} = 9.9$ for engine C. Thus for all of the OCR engines tested, the mean square prediction error is less than 10%.
4. Model scatterplots demonstrate the linearity between the regression model prediction and the actual OCR values. One can also observe that most outliers occur at the lower rates, where text images contain a lot of noise and character recognition rates become less predictable or consistent. For the pages with rates above 60%, regression accuracy increases, resulting in even lower MSE values compared to the overall regression.
5. Both residual histograms and normal P-P plots for each OCR engine suggest that the regression residual distribution approaches the normal, with most residual values close to 0 and only a few less probable outliers.

Thus, the statistical data indicate that the model's predictions are highly correlated with the actual OCR rate for the particular engine. Because the model selection procedure ensured that each of these models provides the best OCR prediction for the given model size and cannot be modified without increasing the mean square error MSE, we therefore obtained a concise and highly significant predictive model for each OCR engine in question.

Apart from this regression analysis, a few descriptive conclusions can be drawn:

1. Each engine has a different set of best predictive features. Because all features in Table 2 are significant, we can analyze the sensitivity of a particular OCR engine to different text quality factors. For instance, engine D is the only one affected by the height of white connected components in the text line regions (WLLCH). High values of this parameter can correspond to a large number of characters broken along the vertical axis, thereby significantly decreasing engine D's recognition rate. Therefore, the knowledge about the optimal predictor set can be used to judge how a particular OCR engine is sensitive to specific text degradation features, and that in turn may suggest ways to improve that engine.
2. All OCR engines exhibit very similar predictability, which can be seen in Table 2 and Figure 5. Note that we used only the most general text quality factors, such as noise and size variation, and did not make any assumptions about character shapes, fonts, or languages.[8] This method leads to an interesting conclusion: Essentially, the character-based OCR rate can be predicted with low MSE values without using character-, language-, or context-specific information and assuming only a certain consistency in text parameters.

Nonlinear Regression Models

Even optimally chosen linear regression models may not be able to account for all possible functionality in the OCR rate prediction. Therefore, nonlinear regression models can also be used to achieve better OCR prediction results by taking advantage of more complicated dependencies. This feature enables us to increase prediction quality without increasing the set of the document quality parameters. Because each parameter has a certain computational cost, the use of nonlinear models can be beneficial in developing a fast document rating system.

Several alternatives for the nonlinear regression analysis were considered as possible improvements to the linear regression models. The rationale for this can be seen in Figure 6 (left), which shows the error $e = D - \hat{D}$ in the engine D fifth-order linear regression $\hat{D}$ versus the engine's actual OCR values D. One can observe that low OCR values tend to be overestimated by the linear regression model, producing mostly negative residual values, and increased OCR rates lead to less negative residuals. Therefore, linear regression was not able to account for all existing dependencies

[8]Except that the languages use English-like alphabets.

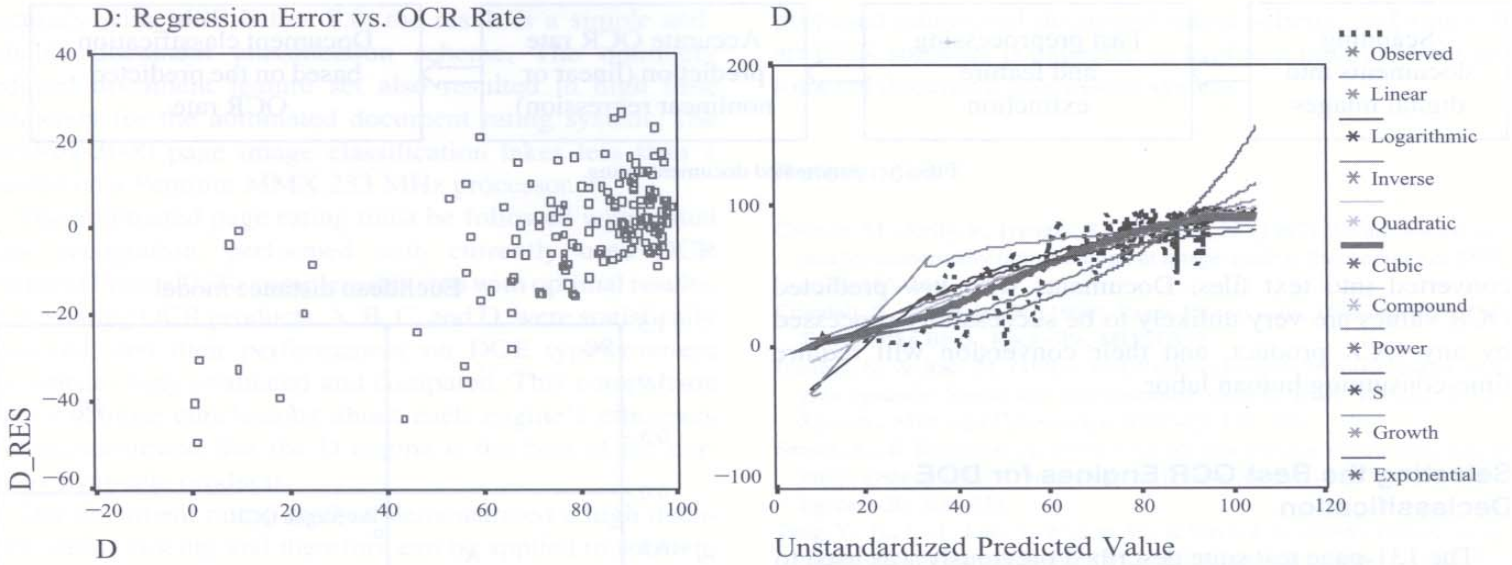


FIG. 6. Engine D nonlinear models.

between the five-feature set values and the actual engine D recognition rate.

To remove these dependencies, we applied the SPSS® curve estimation tool to fit D with various nonlinear functions $D = f(\hat{D})$, as shown in Figure 6 (right). It was found that the cubic model of the form

$$D = \lambda_0 + \lambda_1 \hat{D} + \lambda_2 \hat{D}^2 + \lambda_3 \hat{D}^3,$$

represented by the thick curve in Figure 6 (right), achieved the best performance over the linear regression results. This decreased the regression MSE from 9.36 (as for the fifth-order linear regression) to 8.73, with optimal λ values determined as

$$\lambda_0 = 2.7218,\ \lambda_1 = -0.02526,$$
$$\lambda_2 = 0.0323,\ \lambda_3 = -0.0002.$$

The second alternative was the use of quadratic regression to predict the same actual engine D OCR rate D as

$$D = \alpha_0 + \sum_k \alpha_k p_k + \sum_{i,j} \alpha_{i,j} p_i p_j + e$$

where the p_i's are the document quality parameters used in the linear regression prediction and α_k, $\alpha_{i,j}$ are optimally chosen constants. The best $n = 5$ engine D parameters {*BBLD*, *LCCV*, *RLL*, *TCCV*, *WLLCH*} were chosen from Table 2 as the p_i's. The $\sum_{i,j} \alpha_{i,j} p_i p_j$ term adds the cross-products $p_i p_j$ and the quadratic terms p_i^2 to the linear regression, which leads to considerable improvement in the prediction of OCR accuracy rates, as shown in Table 3.

Note that all quadratic regression terms in this model have influential significance levels but essentially are combinations of only the five document quality parameters used in the linear regression. The MSE value produced by this model was as low as 7.75. The same analyses were performed for the remaining three OCR engines, with very similar (below 8.0 MSE) low prediction errors.

Automated Document Rating

The regression analysis and optimal parameter set selection demonstrated a high potential for predicting OCR accuracy rates and are expected to lead to the automated OCR document recognition rating system shown in Figure 7.

Note that MSE values as low as 8–9, achieved with the proposed OCR prediction process, allow one both to predict the OCR rate produced by a specified OCR engine accurately and to estimate the time and cost of the document classification. Documents with high predicted OCR values can be efficiently recognized by OCR software and

TABLE 3. Engine D quadratic regression results.

Regression	Degrees of freedom	Type I sum of squares	*R*-square	*F*-ratio	Prob > *F*
Linear	5	28650	0.7299	95.251	0.0000
Quadratic	5	747.963587	0.0191	2.487	0.0359
Cross-product	10	3477.306326	0.0886	5.780	0.0000
Total Regression	20	32876	0.8375	27.325	0.0000

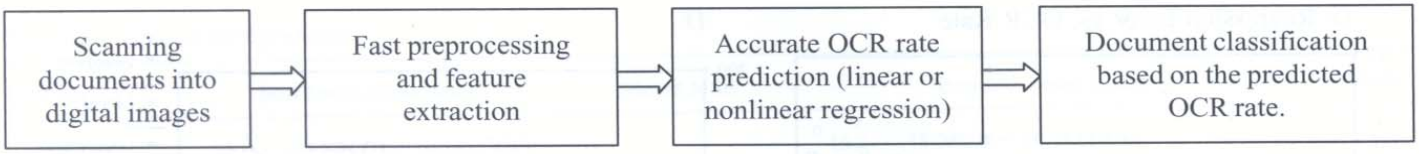


FIG. 7. Automated document rating.

converted into text files. Documents with low predicted OCR values are very unlikely to be successfully processed by any OCR product, and their conversion will require time-consuming human labor.

Selecting the Best OCR Engines for DOE Declassification

The 131-page test suite described previously was used to test the performance of the four OCR engines—A, B, C, and D—to determine which is the most appropriate and efficient for the DOE declassification problem. The four engines were to be compared to each other; that led to the six-pair *t*-test, as shown in Tables A1 and A2 in Appendix B.

1. Paired samples correlations (Table A1) demonstrate that the OCR rates produced by all four products are strongly correlated (the hypothesis of their being uncorrelated has significance levels as low as 0.0001). From these results, engines A and B exhibit the most similar performance, while C and D produce the most dissimilar OCR values.
2. Paired samples means, deviations and standardized errors for each OCR engine are also presented in Table A1. On the basis of the mean OCR rate, the D product delivers the highest average recognition rate (79.2% OCR rate), B (71.6%) is second, A (69.9%) is third, and C produces the lowest OCR values (68.2%). However, these relations between the means are not sufficient to compare the four OCR engines, because they must be related to the possible OCR measurement errors. This comparison was accomplished with the paired *t*-test.
3. The paired samples *t*-test (Table A2) compared all four engines, assuming that the performance of each one can be represented as an accurate recognition rate plus normally distributed error. Because the lowest significance levels in this table correspond to the highest probability that the two engines perform differently, engine D clearly differs from the remaining three products (all three D pairs, and only these, have significance levels below 0.0001). On the other side of the table, D produces the highest average recognition rates and outperforms A by 9.2%, B by 7.5%, and C by 11%. The confidence intervals show that for 99% of the test pages, D is clearly producing better results, outperforming A by at least 6.0%, B by at least 4.1%, and C by at least 7.4%. These numbers demonstrate that D is the most suitable for DOE declassification, producing results that are significantly better than those of the other three OCR products.
4. For the remaining three engines, only B, which has the second best average recognition rate, is different from C, which has the lowest average recognition rate, at the 0.01 significance level.[9] The difference between A and B is less significant than the difference between A and C, so that C is the least suitable product for the DOE declassification problem.
5. The plot in Figure 8 represents the results of the optimal 2-dimensional scaling applied to the four OCR engines and their averaged value. The 2-dimensional coordinates for all five variables on this plot were computed to provide the best Euclidean distance match between the resulting 2-dimensional (on this plot) and the actual the 131-dimensional (in the test observation space) intervariable distances. This graphical result shows that engine D, which has the best OCR rates, is distinctively different from the other three OCR engines and hence is performing at a significantly higher level than the other engines. It also indicates that A and B produce very similar recognition rates, close to the averaged OCR values, and C, which has the lowest OCR rates, is performing at a significantly lower level than the other three engines.

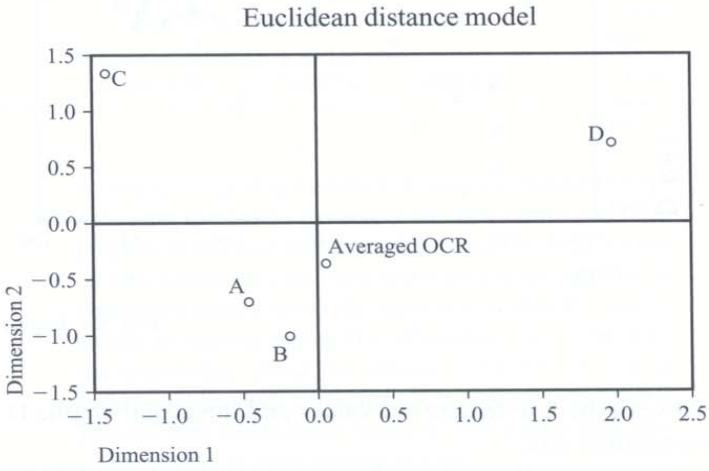


FIG. 8. OCR engines in two-dimensional space.

Conclusions

A new efficient document image technique was proposed. It is based on exploiting the dependencies existing between the character recognition rates produced by currently used OCR engines and optimally chosen image quality factors. The analysis of these dependencies with linear and nonlinear regression tools demonstrates that only five to six document quality parameters are sufficient to achieve OCR prediction

[9]That is, the hypothesis that C performs as well as B may be accepted with only 1% probability.

accuracy with MSE below 8.0; the result is a simple and reliable document classification scheme. The optimally reduced document feature set also resulted in high time efficiency for the automated document rating system: The 2000-by-3000 page image classification takes less than 1 second on a Pentium MMX 233 MHz processor.

The automated page rating must be followed with actual page recognition, performed with currently used OCR engines (Figure 7). To complete this step with optimal results, four existing OCR products, A, B, C, and D, were statistically analyzed, and their performances on DOE typewriter-era documents were evaluated and compared. This comparison led to definite conclusions about each engine's efficiency and demonstrated that the D engine is the best of the currently available products.

Our document rating method demonstrated a high document rating fidelity and therefore can be applied to substantially reduce the costs and labor currently associated with large-scale document processing. The combination of the proposed automated document rating scheme and state-of-art OCR software is expected to lead to a time-efficient automated document processing system.

References

Cannon, M., Kelly, P., Iyengar, S., & Brener, N. (1997). An automated system for numerically rating document image quality. Proceedings of SPIE, 3027, 161–168.

Crandall, S., & Amsler, R. (1996). Pages for test suite, Dyncorp/IRG, 3204 Tower Oaks Blvd., Rockville, MD 20852.

Iyengar, S., & Rao, M. (1983). Statistical techniques in modeling of complex systems: Single and multiresponse models. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13, 175–188.

Senior, A., & Robinson, A. (1998). An off-line cursive handwriting recognition system. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 309–321.

Tang, Y., Tu, L.-T., Lee, S.-W., Lin, W., & Shyu, I.-S. (1998). Offline recognition of Chinese handwriting by multifeature and multilevel classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 556–561.

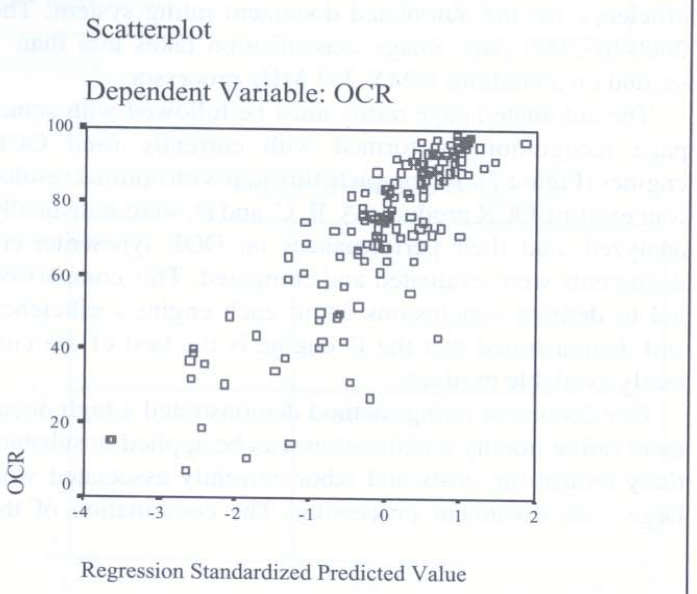
Appendix A: Linear Prediction for OCR Engines: Best Fifth-Order Models

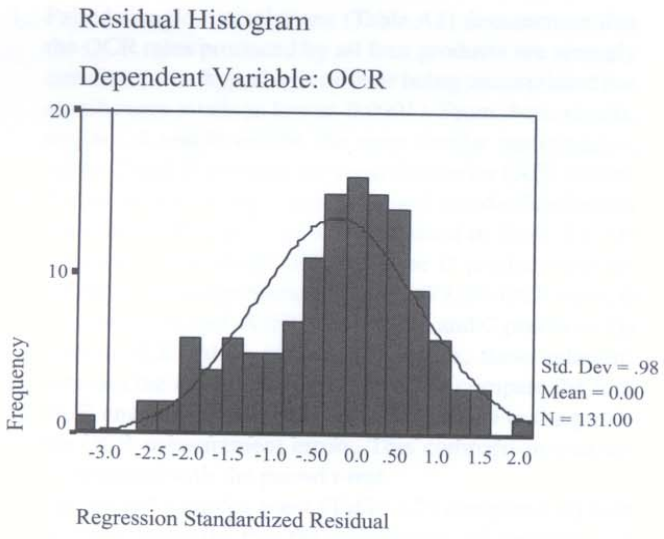
ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	42621.353	5	8524.271	115.235	.000
	Residual	9246.580	125	65.614		
	Total	51867.933	130			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	50.029	4.458		11.221	.000
	BBLD	-123.952	14.736	-.372	-8.412	.000
	CCH	1.122	.171	.321	6.576	.000
	LCCV	.223	.022	1.398	10.126	.000
	TCCV	-.380	.030	-1.801	-12.809	.000
	RLL	18.354	5.787	.169	3.172	.002





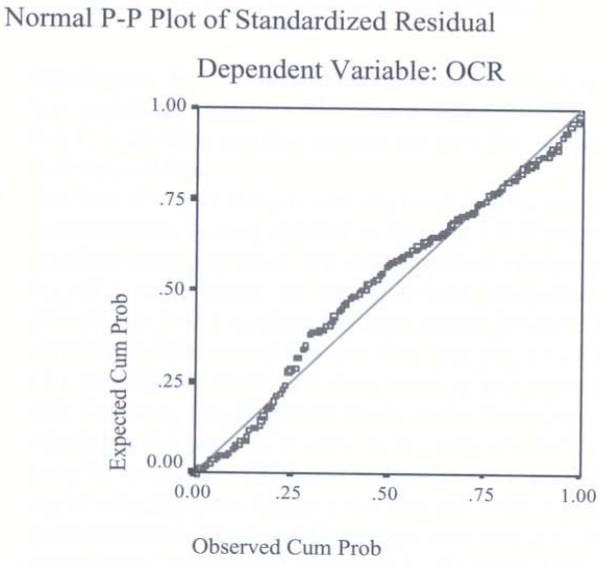


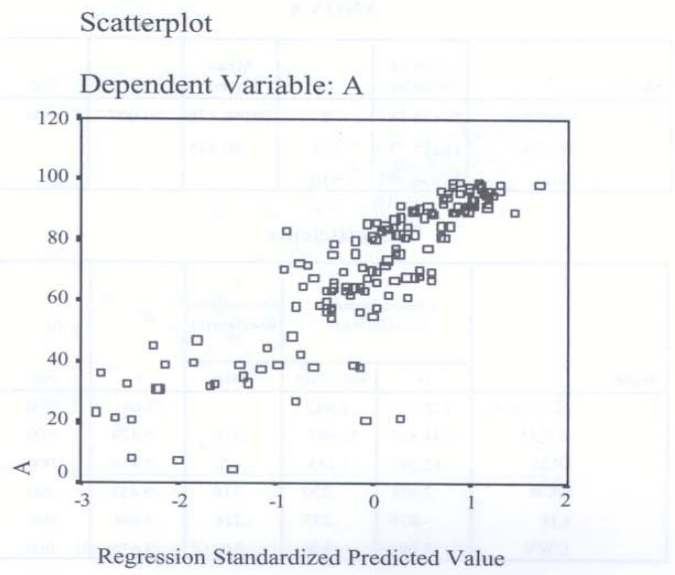
FIG. A1. Averaged OCR regression.

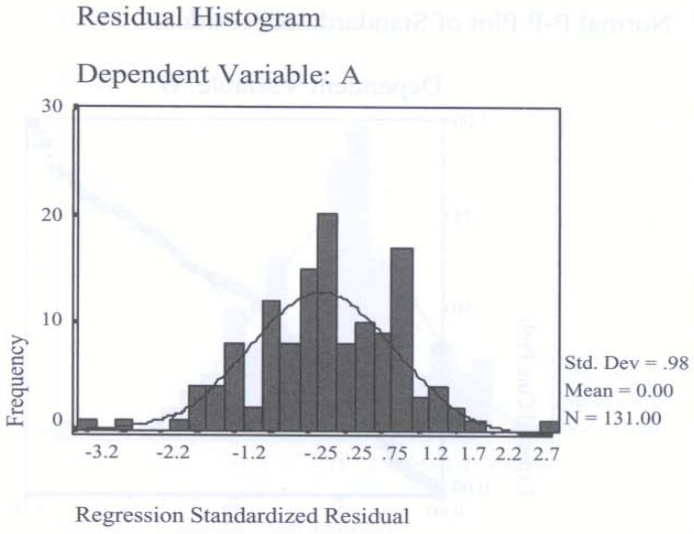
ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	51075.113	5	10215.023	111.761	.000
	Residual	11425.059	125	72.937		
	Total	62500.172	130			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	46.189	4.958		9.316	.000
	RLL	48.001	7.626	.403	6.295	.000
	BLD	-68.251	11.853	-.366	-5.758	.000
	LCCH	1.230	.279	.277	4.406	.000
	LCCV	.164	.028	.936	5.795	.000
	TCCV	-.331	.040	-1.427	-8.280	.000





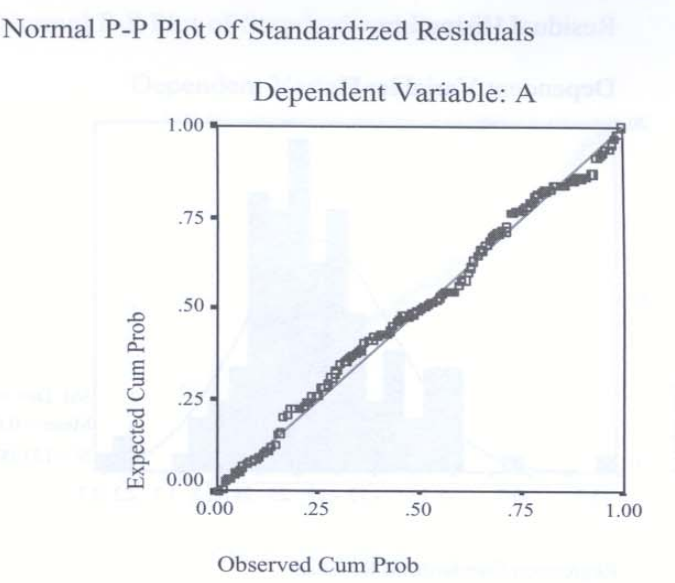


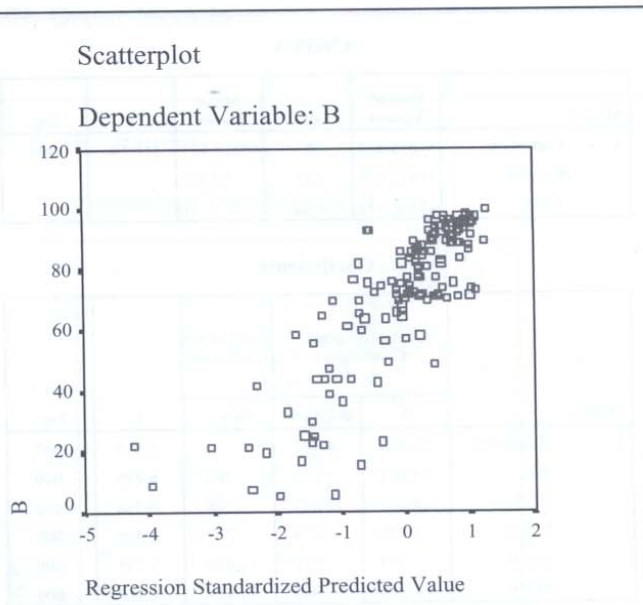
FIG. A2. Engine A regression.

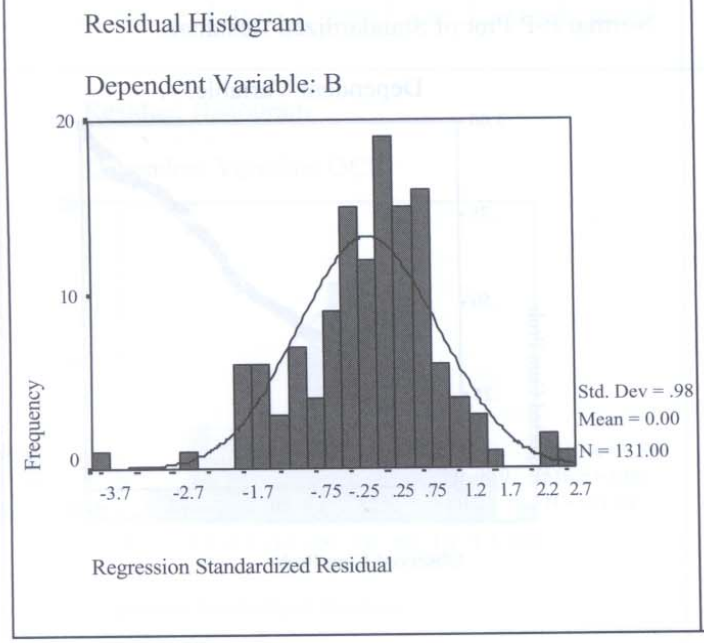
ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	50840.881	5	10168.176	70.007	.000
	Residual	18155.710	125	96.825		
	Total	68996.591	130			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	122.712	8.982		13.661	.000
	BBLD	-144.422	22.463	-.376	-6.429	.000
	BLD	-82.297	15.135	-.420	-5.438	.000
	CCH	2.078	.220	.516	9.458	.000
	CH	-.875	.237	-.216	-3.686	.000
	CWV	-3.085	.840	-.218	-3.674	.000





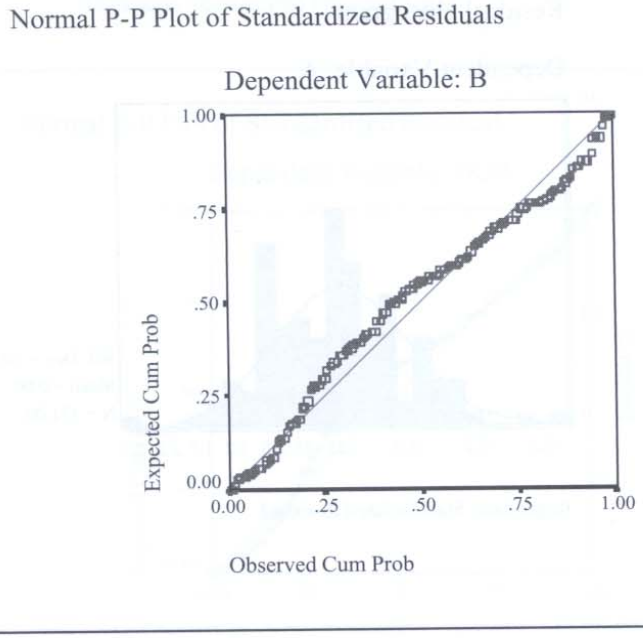


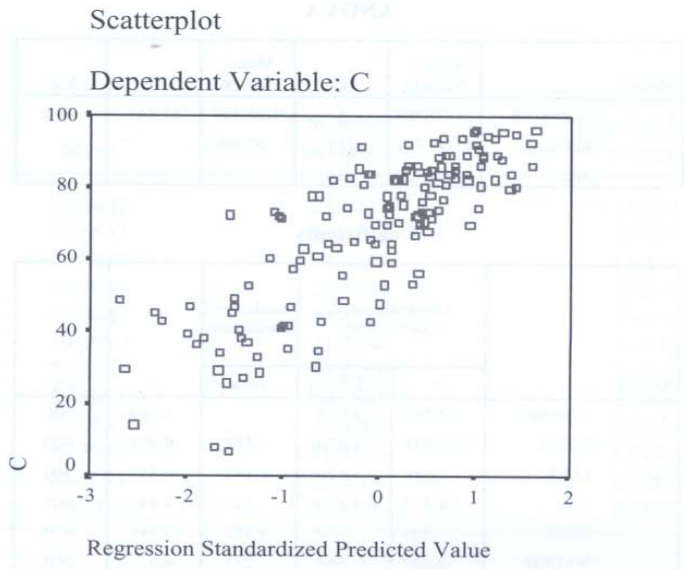
FIG. A3. Engine B regression.

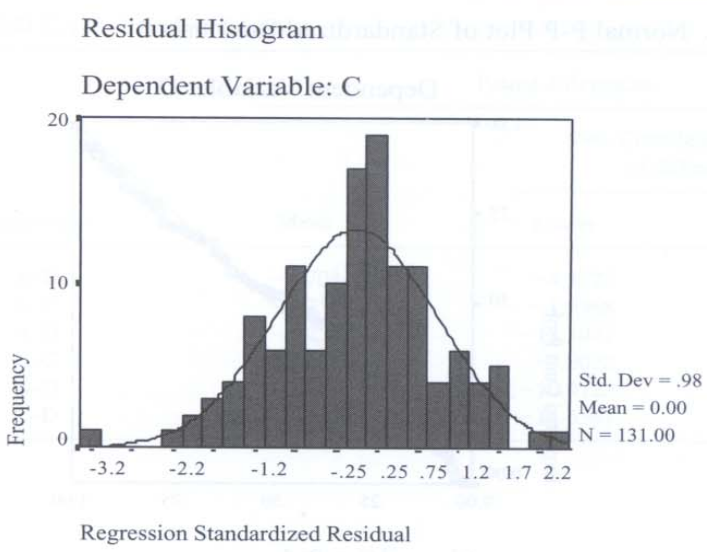
ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	34461.589	5	6892.318	56.092	.000
	Residual	15359.484	125	98.601		
	Total	49821.073	130			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-13.722	13.983		-.981	.328
	BLD	-81.256	11.262	-.488	-7.215	.000
	CCH	1.453	.220	.424	6.591	.000
	CW	-2.730	.521	-.345	-5.245	.000
	RLL	136.281	15.725	1.281	8.666	.000
	RSS	135.680	26.331	.699	5.153	.000





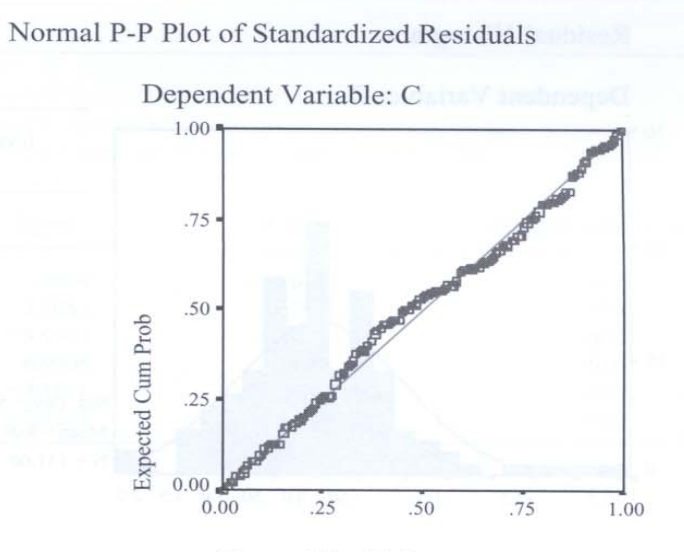


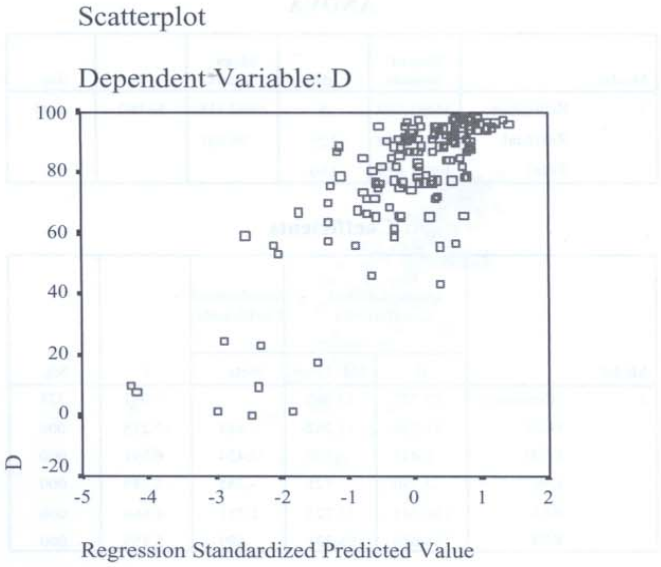
FIG. A4. Engine C regression.

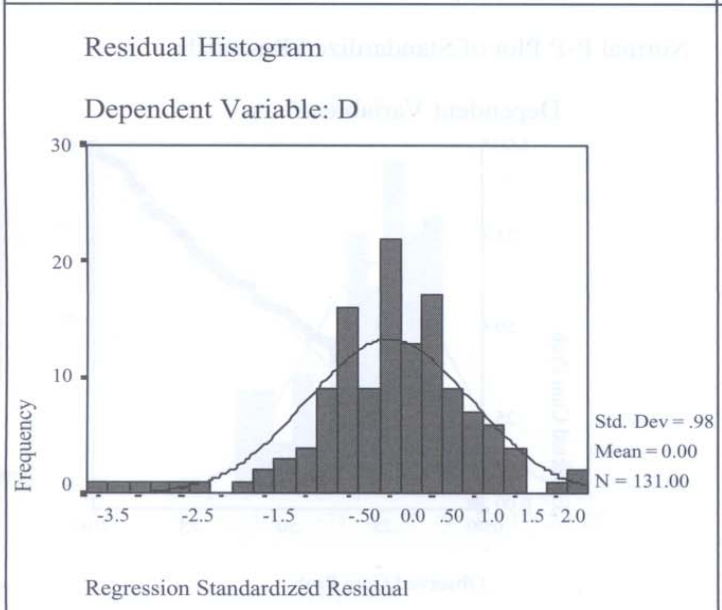
ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36177.988	5	7235.598	57.445	.000
	Residual	15744.739	125	87.609		
	Total	51922.727	130			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	92.763	9.614		9.649	.000
	BBLD	-187.611	21.070	-.562	-8.904	.000
	LCCV	.188	.029	1.182	6.558	.000
	RLL	24.477	7.112	.225	3.442	.001
	TCCV	-.300	.039	-1.422	-7.741	.000
	WLCCH	-2.693	.668	-.233	-4.029	.000





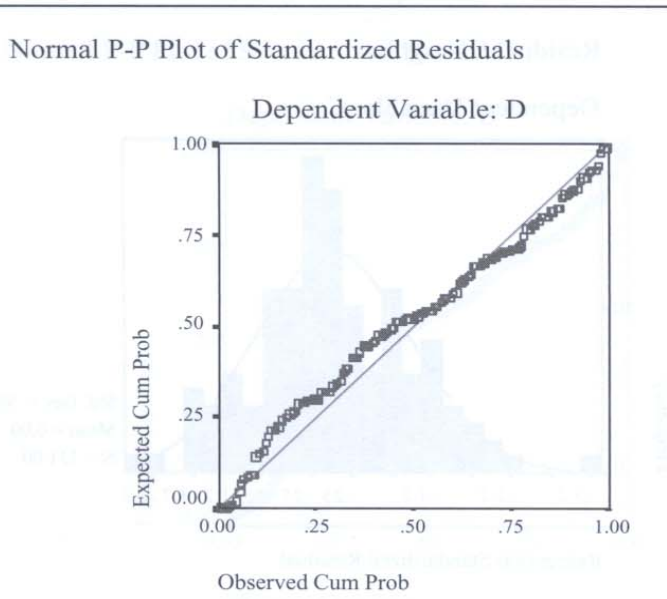


FIG. A5. Engine D regression.

Appendix B: Comparing OCR Engines

TABLE A1. Paired samples statistics.

Engine pair	Correlation	Correlation significance	Mean	Standard deviation	Standard error of mean
A–B	.870	.000	69.9449	23.4927	1.9577
			71.6533	24.4268	2.0356
A–C	.764	.000	69.9449	23.4927	1.9577
			68.1891	20.7201	1.7267
A–D	.784	.000	69.9449	23.4927	1.9577
			79.1970	21.7580	1.8132
B–C	.760	.000	71.6533	24.4268	2.0356
			68.1891	20.7201	1.7267
B–D	.773	.000	71.6533	24.4268	2.0356
			79.1970	21.7580	1.8132
C–D	.690	.000	68.1891	20.7201	1.7267
			79.1970	21.7580	1.8132

TABLE A2. Paired samples test.

	Paired differences				
		99% Confidence interval of difference			
Engine pair	Mean	Lower	Upper	*T*-test	Standard error of mean
A–B	−1.7084	−4.3772	.9604	−1.671	.097
A–C	1.7558	−1.5968	5.1084	1.367	.174
A–D	−9.2521	−12.5072	−5.9969	−7.420	.000
B–C	3.4642	−0.0022	6.9506	2.594	.010
B–D	−7.5437	−10.9723	−4.1151	−5.744	.000
C–D	−11.0079	−14.6547	−7.3611	−7.880	.000